Comparative Analysis of Nonlinear Models Developed using Machine Learning Algorithms

MAJA ROŽMAN[1], ALEN KIŠIĆ[2], DIJANA OREŠKI[3]
[1]Faculty of Economics and Business,
University of Maribor,
Razlagova ulica 14, 2000 Maribor,
SLOVENIA

[2]Vern University,
Palmotićeva 82/1, 10 000 Zagreb,
CROATIA

[3]Faculty of Organization and Informatics,
University of Zagreb,
Pavlinska 2, Varazdin,
CROATIA

Abstract: - Machine learning algorithms are increasingly used in a vast spectrum of domains where statistical approaches were previously used. Algorithms such as artificial neural networks, classification, regression trees, or support vector machines provide various advantages over traditional linear regression or discriminant analysis. Advantages such as flexibility, scalability, and improved accuracy in dealing with diverse data types, nonlinear problems, and dimensionality reduction, compared to traditional statistical methods are empirically demonstrated in many previous research papers. In this paper, two machine learning algorithms are compared with one statistical method on highly nonlinear data. Results indicate a high level of effectiveness for machine learning algorithms when dealing with nonlinearity.

Key-Words: - Machine learning, decision tree algorithm, artificial neural network, predictive models, data characteristics, nonlinear data, artificial intelligence.

Received: July 18, 2023. Revised: April 21, 2024. Accepted: May 16, 2024. Published: June 20, 2024.

1 Introduction

This paper is written under the project SIMON: *Intelligent system for automatic selection of the machine learning algorithms in the social sciences*. The main objective of the SIMON project is to develop an intelligent system that can automatically recommend machine learning algorithms in the social sciences that work better on a particular data set while taking into consideration the data properties of educational and business datasets. Comparative analysis of various machine learning algorithms across a huge number of datasets is a key component of the project's research. In this paper, the focus is on educational data and two groups of machine learning algorithms, machine learning based on error and machine learning based on information. Those algorithms are compared with an advanced statistical approach of discriminant analysis on the dataset which is characterized by high level of nonlinearity. Data properties are measured through meta-features. There are various categories of meta-features. In this research, meta-features that explain data linearity are employed.

Artificial intelligence and machine learning development transformed all aspects of our lives. Education is a field that is continuously adapting to new technologies. Development and application of artificial intelligence and machine learning in education have opened new research paths and possibilities for predicting student performance, mostly in creating personalized learning approaches and understanding the factors contributing to educational success. The combination of powerful machine learning algorithms and huge amounts of data created and stored from learning management systems (LMS) leads to significant scientific achievements. This paper investigates the predictive power of two machine learning algorithms of

different approaches to learning and one advanced statistical method for LMS data analysis and the development of student success predictive models which serve as the basis for personalized intelligent systems in education. The research presented here aims to evaluate the effectiveness and application of machine learning algorithms on nonlinear data.

The paper is structured as follows. Section two reviews previous related research papers. Section three provides a description of data and a brief overview of machine learning algorithms. Section four gives insights into research results. Section five concludes the paper and gives guidelines for future research.

2 Theoretical Backgrounds

2.1 Machine Learning Algorithms for Nonlinear Predictive Models' Development

Machine learning algorithms are playing an increasingly important role in analyzing datasets that are characterized by nonlinearity. By employing diverse approaches to the learning phase and development of descriptive and predictive models, machine learning algorithms strive to enhance accuracy and minimize biases. These algorithms are particularly beneficial in the educational context, as they can provide more reliable insights regarding student performance and instructional strategies, especially when the data are for technology-enhanced platforms and/or learning management systems. Moreover, machine learning algorithms contribute to the identification of patterns and trends within educational data, enabling educators and policymakers to make data-driven decisions to improve teaching and learning outcomes. However, the evaluation of machine learning algorithms on social science datasets also presents some challenges. It is crucial to examine the interpretability and explainability of machine learning algorithms to facilitate understanding of models. Social science domains are characterized by complexity and nonlinearity. Evaluation of machine learning algorithms on educational datasets has the potential to revolutionize the education sector, but it requires careful consideration of various factors to maximize their impact. Hereinafter, an overview of the machine learning algorithms is provided on educational datasets.

2.2 Machine Learning for Predictive Models' Development

Machine learning algorithms used on learning management system data for predicting student success include bagging, boosting, stacking, and voting, [1]. The proposed models in the study, [2], integrate five traditional machine learning algorithms (DT, RF, GBT, NB, and KNN) with ensemble techniques, resulting in improved prediction performance. Another study, [3], explores the use of bagging algorithms like random forest and boosting algorithms like adaptive boosting, stochastic gradient boosting, and extreme gradient boosting for predicting student performance. Additionally, the study, [4], compares and analyzes five ensemble classifiers, including bagging decision trees, for modeling student behavior from e-learning data. Authors in [5], performed a thorough exploration and analysis of two educational datasets. Proposed ensemble models achieve high accuracy and low false positive rates.

There are various papers using machine learning algorithms for student performance prediction, such as [6] and [7]. E.g. [7], tested several models for predicting student success and the support vector machines algorithm achieved the best results with a prediction rate of 87.32%. Project SIMON researchers also contributed to the research topic by examining machine learning algorithms applications in social sciences in general [8], focusing on educational data and developing predictive models by using different approaches such as machine learning based on probability or by comparing various machine learning approaches, [9].

This research takes a step forward by comparing machine learning algorithms with algorithms based on statistical learning, and discriminant analysis.

3 Research Methodology

3.1 Learning Management System Data

Educational datasets include a wide range of information such as demographic details, academic performance, behavioral patterns, and more. These datasets provide a comprehensive view of the student learning process, making them a basis for applying machine learning algorithms. Nowadays, LMS data are the primary source due to the large level of LMS usage in education.

The dataset used in this research is extracted from Knowledge Discovery in Data course, taught at the Faculty of Organization and Informatics, University of Zagreb. The course was offered as an

elective at the undergraduate level of study program Information and Business Systems. The course was taught as a blended learning approach, combining traditional classroom-based lectures and laboratory exercises with the various tasks and activities at the Learning Management System Moodle.

Data extraction was carried out from the Moodle platform for two separate student generations, thereby making a sample size of 83 students. Raw data about students' activities were extracted and measured by the number of students' logs to specific resources and activities. These included files, forums, student reports, folders, choices, file submissions, overview reports, pages, systems, tests, and assignments.

The final course grade was incorporated into the dataset as a dependent variable, thus enabling the application of supervised machine learning algorithms and the development of predictive models of student success.

3.2 Machine Learning Algorithms

There are numerous research studies investigating artificial neural networks and decision tree usage for predictive models in education to predict student performance, dropout rates, and misconduct locations, with varying levels of accuracy.

Artificial neural networks and decision trees were used to predict the performance of students in a computer science course at Al-Muthanna University, with a classification accuracy of 77.04%, [10].

Both Decision Trees and Artificial Neural Networks were used to develop classification models and generate rules to classify and predict students' behavior and the location of misconduct on college campuses, [11].

Artificial Neural Network algorithms and Decision Tree algorithms were used for constructing a prediction model of student achievement in business computer disciplines at the School of Information and Communication Technology, the University of Phayao, [12].

Artificial neural networks are error-based machine learning approach that learns by adjusting weights between neurons and thus minimizing the error of the model. The whole idea is based on the biological neurons and the way they function.

Decision trees are information-based machine learning approaches which learn by identifying the most informative variables from the data set and constructing a decision tree model by using the most informative variables.

4 Research Results

There are numerous approaches to evaluate and test predictive model accuracy. In this case, k-fold cross-validation is used. Using k-fold cross-validation, the data set is divided into k subgroups. One of the k subsets is always the test set, while the other k-1 subsets are always the training set. In this study, ten folds are employed.

Table 1. Predictive models' accuracy

Algorithm	Accuracy
Artificial neural network	79.83 %
Decision tree	78.25 %
Discriminant analysis	71.09 %

Using the performance metric from Table 1, we may draw several inferences. It is to be: can the results be generalized, or are they the results of chance? Determining how accurately evaluation measures reflect classifier behavior is the aim of statistical significance testing. We tested the algorithms on one domain and compared them using two matched sampling t-tests. At the significance threshold of 0.05, the mean difference's significance is examined. The assumption in Table 2 is that there is no difference in the mean values of algorithm performances, and this is done to see if we can reject.

Table 2. T-test results

Hypothesis	Model	T-test
H0: Artificial neural network = discriminant analysis	Artificial neural network	0.004
	Discriminant analysis	
H0: Decision tree = discriminant analysis	Decision tree	0.007
	Discriminant analysis	
H0: Artificial neural network = Decision tree	Artificial neural network	0.05
	Decision tree	

As seen in Table 2 there are statistically significant differences in the performances of the two artificial neural networks and discriminant analysis, as well as in decision tree performances when comparing it with discriminant analysis. However, there are no differences in performances between two machine learning approaches: artificial neural networks and decision trees. Results of statistical testing indicate the superiority of machine

learning approaches over statistical learning approach to developing predictive models.

5 Conclusion

In this paper, we have proposed two machine learning-based student predictive models and one statistical learning-based student predictive model. The proposed model adopts two different approaches to machine learning-based development of predictive models: machine learning based on error (artificial neural network) and machine learning based on information (decision tree algorithm). The results of the performance evaluation reveal there are statistically significant differences between machine learning and statistical learning approaches, but there are no statistically significant differences between the two different machine learning approaches.

This paper gives two scientific contributions: i) in the field of machine learning, by investigating how different machine learning approaches handle educational LMS data, (ii)) in the field of statistical learning, by investigating how handles educational LMS data (iii) in student predictive models, by comparing different machine and statistical learning approaches and demonstrating which one achieves the best predictive model in this domain.

There are several limitations of the research presented here. First, only one dataset is used in algorithm comparison. In future research, we will upgrade several datasets including several courses at several study programmes and different faculties and different countries. Also, the LMS data will be subjected to various machine learning algorithms, and their performances will be compared to determine the results.

Findings from this research could help to tailor teaching and learning strategies, particularly in virtual learning environments.

Declaration of Generative AI and AI-assisted technologies in the Writing Process

During the preparation of this work, the authors used Paperpal to improve the language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References:

[1] Saleem, F., Ullah, Z., Fakieh, B., & Kateb, F. (2021). Intelligent decision support system for predicting student's E-learning performance using ensemble machine learning. Mathematics, 9(17), 2078, pp. 1-22, https://doi.org/10.3390/math9172078.

[2] Kumar, M., & Jeet Singh, A. (2022, September). Process-Based Multi-level Homogeneous Ensemble Predictive Model for Analysing Student's Academic Performance. *In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022*, Vol. 1 (pp. 139-159). Singapore: Springer Nature Singapore.

[3] Lenin, T., & Chandrasekaran, N. (2021). Learning from Imbalanced Educational Data Using Ensemble Machine Learning Algorithms. *Webology*, 18(SI01), 183-195.

[4] Hassan, H., Ahmad, N. B., & Sallehuddin, R. (2021). An Empirical Study to Improve Multiclass Classification Using Hybrid Ensemble Approach for Students' Performance Prediction. *In Computational Science and Technology: 7th ICCST 202*0, Pattaya, Thailand, 29–30 August, 2020 (pp. 551-561). Springer Singapore.

[5] Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, 200, 105992, https://doi.org/10.48550/arXiv.2005.06647.

[6] Ahamad, M., & Ahmad, N. (2021). Students' knowledge assessment using the ensemble methods. *International Journal of Information Technology*, 13(3), 1025-1032.

[7] Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting student success using big data and machine learning algorithms. *International Journal of Emerging Technologies in Learning*, 17(12), 236.

[8] Oreski, D. (2023). Application of Machine Learning Methods for Data Analytics in Social Sciences. *WSEAS Transactions on Systems*, 22, 69-72, https://doi.org/10.37394/23202.2023.22.8.

[9] Oreški, D., & Hajdin, G. (2019). Development and comparison of predictive models based on learning management system data. *WSEAS Transactions on Information Science and Applications*, 16, 192-201.

[10] Altabrawee, H., Ali, O., & Ajmi, S. (2019). Predicting Students' Performance Using Machine Learning Techniques. *Journal of University of Babylon for Pure and Applied Sciences*, 27(1), 194-205, https://doi.org/10.29196/JUBPAS.V27I1.2108

[11] Blasi, A., & Alsuwaiket, M. (2020). Analysis of Students' Misconducts in Higher Education using Decision Tree and ANN Algorithms. *Engineering, Technology & Applied Science Research*, 10, 6510-6514, https://doi.org/10.48084/etasr.3927.

[12] Nuankaew, P., Nuankaew, W., Teeraputon, D., Phanniphong, K., & Bussaman, S. (2020). Prediction Model of Student Achievement in Business Computer Disciplines. *Int. J. Emerg. Technol. Learn.*, 15, 160-181, https://doi.org/10.3991/ijet.v15i20.15273.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This paper is supported by Croatian science foundation under the project SIMON: Intelligent system for automatic selection of machine learning algorithms in social sciences, UIP-2020-02-6312.

Conflict of Interest

The authors have no conflicts of interest to declare.